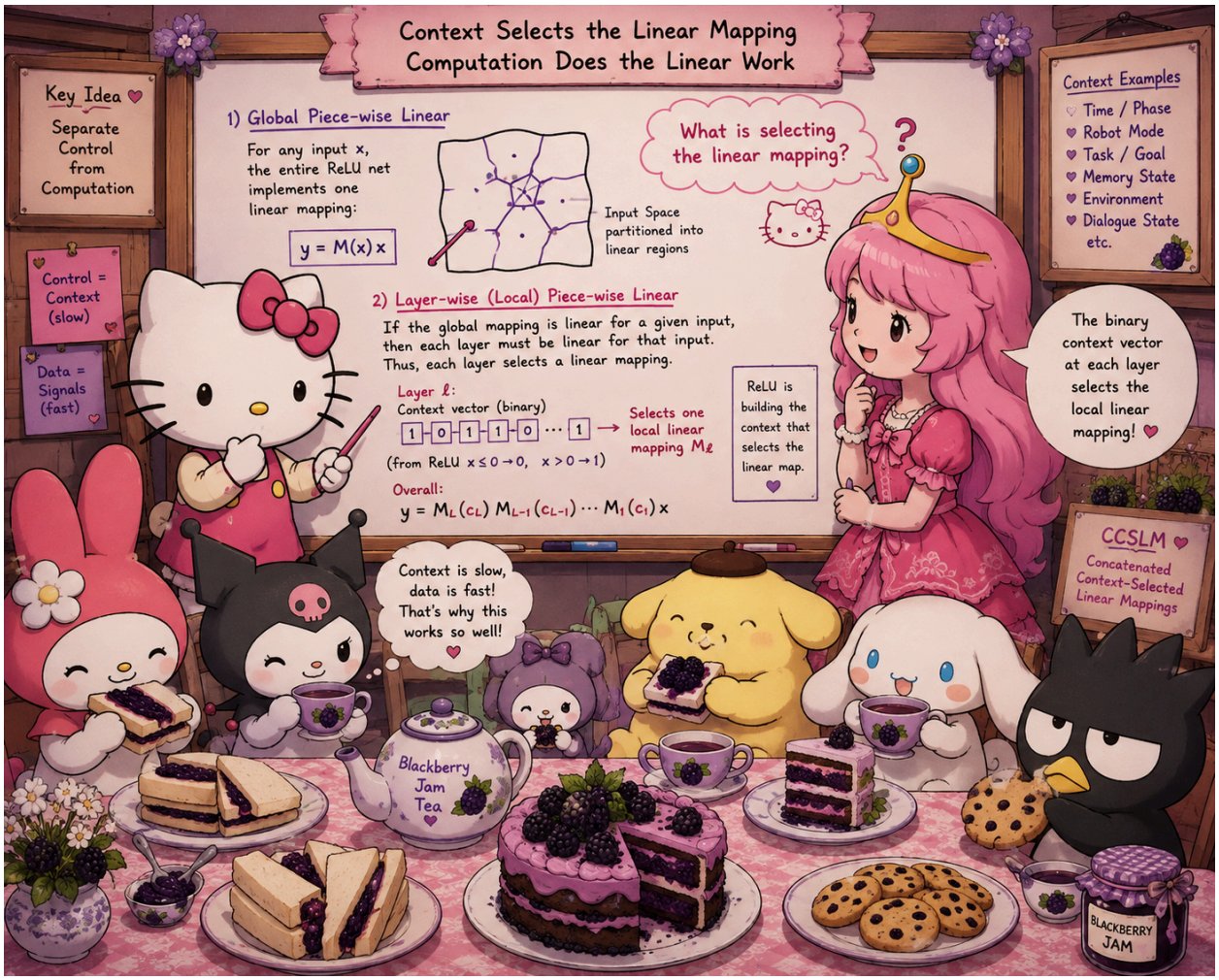
Context Selects the Linear Mapping
Computation Does the Linear Work
Key Idea
Separate Control from Computation
1) Global Piece-wise Linear
For any input x, the entire ReLU net implements one linear mapping:
y = M(x) x
Input Space partitioned into linear regions
What is selecting the linear mapping?
Context Examples
Time / Phase
Robot Mode
Task / Goal
Memory State
Environment
Dialogue State
etc.
Control = Context (slow)
Data = Signals (fast)
2) Layer-wise (Local) Piece-wise Linear
If the global mapping is linear for a given input, then each layer must be linear for that input. Thus, each layer selects a linear mapping.
Layer ℓ:
Context vector (binary)
1 0 1 1 0 ⋯ 1
Selects one local linear mapping Mℓ
(from ReLU x ≤ 0 → 0, x > 0 → 1)
Overall:
y = M_L(c_L) M_{L-1}(c_{L-1}) ⋯ M_1(c_1) x
ReLU is building the context that selects the linear map.
The binary context vector at each layer selects the local linear mapping!
CCSLM
Concatenated Context-Selected Linear Mappings
Context is slow, data is fast! That's why this works so well!
Blackberry Jam Tea
BLACKBERRY JAM

ReLU neural networks are known to be globally piece-wise linear, its layers must also be piece-wise linear.